Tulu 3 Latency Scaling Across Context Lengths in Complex Reasoning Tasks

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the latency scaling curve of Tulu 3 vary across different context lengths during complex reasoning tasks compared to base Llama 3.1 models. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research question: How does the latency scaling curve of Tulu 3 vary across different context lengths during complex reasoning tasks compared to base Llama 3.1 models?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

14 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Gemma-2-9B-it achieves 65.4% accuracy on MMLU and 68.1% on Mobile-MMLU.	×	0.05
Qwen2.5-7B-instruct, Llama-3.1-8B-instruct, Qwen2.5-3B-instruct, Phi-3.5-mini-instruct, Llama-3.2-3B-instruct, Gemma-2-2	×	0.06
The performance spread on MMLU ranges from 45.9% to 71.8%.	×	0.02
The performance spread on MMLU-Pro ranges from 7.5% to 36.5%.	×	0.07
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.06
Qwen2.5-3B-Instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
Llama-3.2-3B-Instruct scores 50.2% accuracy on Mobile-MMLU.	×	0.05
The mean accuracy across models on Mobile-MMLU is 46.84%.	×	0.06
Phi-3.5-mini-instruct achieves 63.7% accuracy on Mobile-MMLU.	×	0.05

References

- http://arxiv.org/abs/2411.15124v5
- http://arxiv.org/abs/2604.05114v1
- http://arxiv.org/abs/2503.20786v1